Chapter I-6. Subjective Probability and Expert Elicitation

Key Concepts

Many situations arise in the course of performing risk analyses for dams where insufficient statistical information exists and models for calculating probabilities simply don't exist. In order to make quantitative risk estimates, it then becomes necessary to judge the likelihood of various events or conditions. This can be done by asking someone who should know how to judge the relevant event probabilities for the failure mode being discussed. Probabilities are then estimated or assigned using subjective, degree-of-belief probability methods. This may also involve the use of techniques that can be used to infer a probability (verbal transformation, betting, probability wheels). It is best if more than one expert is involved in making the estimates. Subjective estimating is therefore typically done in a team setting where "synergy" enhances and draws out the breadth of experience brought to the table by a group of individuals qualified to make the estimates. Team members can enter into discussions that will allow the group as a whole to arrive at a more comprehensive estimate than each individual could on their own.

A subjective probability estimate is the numerical value or range of values judged to be believable based upon the available evidence. The quantity to be assessed must be clearly defined so that each person has a common understanding. When soliciting judgmental probabilities, there are no 'right' answers. In some cases, asking the experts to argue the opposite point can promote further discussion and understanding of the issue. Anything that promotes hard thinking and insight helps.

There is much that could be said about subjective probability and degree of belief. In fact Steve Vick has written an excellent book, largely on this subject, entitled "Degrees of Belief, Subjective Probability and Engineering Judgment" (Vick, 2002), which is suggested reading for anyone interested in more information on this subject. A few of the more relevant concepts are outlined below. Some commonly-used terms are defined in a glossary at the end of this manual.

Concepts of Probability

Consider the following propositions:

- Flip a coin. It will come up heads.
- The 1,000th digit right of the decimal point in pi is a 7
- There are more than 10,000 telephones in Addis Ababa, Ethiopia
- The 20th president of the U.S. was a Republican
- The U.S. president elected in 2016 will be a Democrat
- The numbers 55, 58, 62, 53, and 52 were obtained by random sampling from a normal distribution with a mean of 50 and a standard deviation of 2

You have information bearing on all these propositions, perhaps as follows:

- A coin is symmetrical
- There are ten possible digits
- Ethiopia is the capitol of a poor, not very highly developed country
- The 20th president was after the Civil War & Republicans did well then
- You have some ideas what could happen between now and the next election
- It doesn't seem likely a random process would select numbers all on one side of the mean, and so far from the mean with such a small standard deviation

On the basis of these and other items, you have opinions about the truth of these propositions. It is the proposition and not the event about which you have an opinion. Uncertainty is a property of your knowledge about these events, and not of the events themselves (von Winterfeldt and Edwards, 1986).

By calculating Annual Failure Probability and Risk as we do, Reclamation asserts that:

- All uncertainties are inherently of the same kind (insufficient knowledge)
- Probabilities are useful numbers with which to measure uncertainties
- Probabilities are, in general, personal degrees of belief about uncertain events

Relative Frequency

Different kinds of observations of the real world are related to the probability numbers we use to represent uncertainty. **Relative frequency** relates to the observed frequency of events or outcomes. Suppose you must decide whether to undergo surgery and the physician says: "My experience with this operation in the past has been excellent. People almost never die on the table, and recovery usually takes a week. In a month you'll be as good as new." How useful would it be to replace the verbal statements of uncertainty with numbers? One could count the number of times the operation has been performed and the number of times someone has died to obtain a relative frequency of failure. But what about the age and physical condition of the patient, the doctor's experience, the nature of potential complications, presence of pathogens and others specific factors?

Consider the following: How many times in a day and times in a year do you drive your car? When was the last time you got into an accident? How would you calculate the probability of getting in an accident each time you drive your car? How about the probability per year of getting in an accident? Is there something fundamentally different about considering a break-down instead of an accident?

The desire to properly condition the counting conflicts with the desire to have a large sample. That is, as the samples are subdivided into smaller and smaller groups, each group contains a smaller number of samples. Combinations (samples that belong to more than one group) make it difficult to decide to which counting category a particular case belongs. The way the counting is conditioned, affects how appropriate it is to apply it to estimating an event probability. Changing conditions with time can also affect relative frequency. It's not perfect, but estimating based on relative frequency is the best we can do in some cases.

Verbal Mapping

Reclamation has adopted a verbal mapping scheme for most of the subjective probability estimates. Subjective probability estimates are typically made to represent the likelihood of each event for a potential failure mode that has been decomposed and modeled in an event tree. The mapping scheme adopted by Reclamation is based primarily on experiments reported by Reagan et al (1989). These experiments show that, within reasonable limits, people are pretty well calibrated and consistent relative to known probabilities, provided they use words that most people would adopt on their own and provided the likelihoods are greater than 0.01. Vick (2002) has summarized those results and proposed a verbal to numerical transformation convention, shown in Table 13-1, along with approximate results of the experiments performed by Reagan et al.

A key finding of the experiments was that people's ability to judge likelihood does not extend very far out on either end of the probability scale (i.e. to more than a couple orders of magnitude as one approaches either zero or one), even when words like "almost impossible" or "almost certain" are used. This is likely due to the fact that most people's experience does not allow them to conceptualize likelihoods at these extreme probabilities, and thus we do not have words that adequately describe them. In addition, it is difficult for people to conceptualize how often events happen at remote probabilities absent actual frequency data.

Table I-6-1. Verbal Transformations proposed by Vick (2002)

Verbal Descriptor	Suggested Probability	Approximate Probability Range from Reagan et al
Virtually Impossible, due to known physical conditions or processes that can be described and specified with almost complete confidence	0.01	0-0.05
Very Unlikely, although the possibility cannot be ruled out	0.1	0.02-0.15
Equally Likely, with no reason to believe that one outcome is more or less likely than the other (when given two outcomes)	0.5	0.45-0.55
Very Likely, but not completely certain	0.9	0.75-0.9
Virtually Certain, due to known physical processes and conditions that can be described and specified with almost complete confidence	0.99	0.9-0.995

Despite the evidence to suggest that these are about the limits that people are well calibrated, Reclamation made the decision to add one more order of magnitude to the upper and lower end of the scale, such that in special situations higher or lower numbers could be provided. However, it was recognized that these numbers should rarely be used, and only in situations where conditions are perceived to be so much more unlikely than the ordinary ranges to which the team has been accustomed to estimating, that an additional order of magnitude is warranted. For example, a team can become calibrated to different levels of likelihood as events are discussed and likelihoods are estimated. After frequent assignments in the range of 0.01 to 0.05, an event is encountered that is perceived to have an even lower likelihood. Then it would be reasonable to assign a value as low as 0.001. This led to the verbal mapping scheme shown in Table 13-2 as that used for the majority of Reclamation risk analyses. However, if at all possible, events as remote as 0.001 should be decomposed into two or more contributing events, which presumably would get the estimating back into the more familiar ranges.

Table I-6-2. Verbal Mapping Scheme Adopted by Reclamation

Descriptor	Assigned Probability
Virtually Certain	0.999
Very Likely	0.99
Likely	0.9
Neutral	0.5
Unlikely	0.1
Very Unlikely	0.01
Virtually Impossible	0.001

Other verbal mapping schemes have been proposed that may not correspond well with the variations described above. For example, Barneich et al (1996) published a scheme for estimating conditional probabilities. This was apparently adapted from the 1984 Military Standard System Safety Program Requirements of the Department of Defense (882B, 1984), which at the time were qualitative. The Department of Defense standard was updated to 882D in 2000 to include numerical values associated with the qualitative descriptions. However, it was noted that the actual descriptions and likelihoods are dependent on the size of the inventory, and that adjustments may be needed. The Barneich descriptors have been modified from those published by the Department of Defense. The scheme represents an "order of magnitude" assessment ranging from 10^{-1} to 10^{-4}. The basis for the numbers assigned to the descriptor categories in this mapping scheme is not clear.

Weighing the Evidence and Making Estimates

Once a potential failure mode has been decomposed, estimates for each event can be made using the verbal mapping scheme described above. This is done similar to the procedure described for failure mode screening (see Section on Potential Failure Mode Identification, Description, and Screening), by listing the adverse factors (factors that make the branch more likely) and favorable factors (factors that make the branch less likely) associated with each branch. These factors are the evidence that is used to reach a likelihood estimate.

Both the strength of the evidence, or how convincingly it supports one position or another (how compelling the factor is or how strongly the reasoning ties the evidence to the conclusion), and the weight of the evidence, or how good the evidence is in terms of quality and quantity (how much confidence there is in the evidence), must be considered. For example, if there are lots of good quality SPT tests, the evidence for a particular representative blow count would have a lot of "weight". If that representative blow count was low (say in the range of 3 to 6), reasoning that liquefaction is likely would have a lot of "strength".

The process of collecting and listing the "pros and cons" by the facilitator typically generates significant discussion, which in turn provides a relative sense of the importance of each factor discussed. Once the facilitator senses discussion has ended, the group is queried for a probability estimate or range of estimates using the verbal descriptors outlined in Table 13-2. It is important to draw out the diverging opinions, as they could represent different views on the topic, and may represent the bounds for the estimates.

After the estimates or range of estimates are made, the key factors leading to that estimate must be captured and documented. For example, if the estimate is a low number on the unlikely side, what are the one to three or so key factors that led to the low estimate? These would be the factors with the highest weight and/or strength. An attempt is made to reach consensus on a best estimate and a range of estimates. However, differences in opinion are beneficial to the process, and if consensus cannot be reached, the diverging views are documented along with their reasoning and the resulting estimates. The opposing estimates are carried through to estimates of annual failure probability and risk, and both sets are reported. The facilitator helps guide the process, on the lookout for biases or adverse group interactions.

Distributions

During an Issue Evaluation Risk Analysis, it is often necessary to estimate probability distributions for nodes on an event tree or other parameters that affect the risk estimates. If done in a team setting using judgmental probabilities, there are typically only a handful of distributions that are useful. The facilitator typically will direct the development of a distribution by asking a series of questions, typically starting with, "What is the lowest you can imagine the likelihood to be, and what is the highest you can imagine the likelihood to be?" Then the question becomes, "is it more likely to be somewhere in between these values?" If the answer is no, then it is equally likely to be anywhere between the two limits, and the uniform distribution, as shown in Figure 13-1, is appropriate. The distribution mean is the average of the maximum and minimum values.

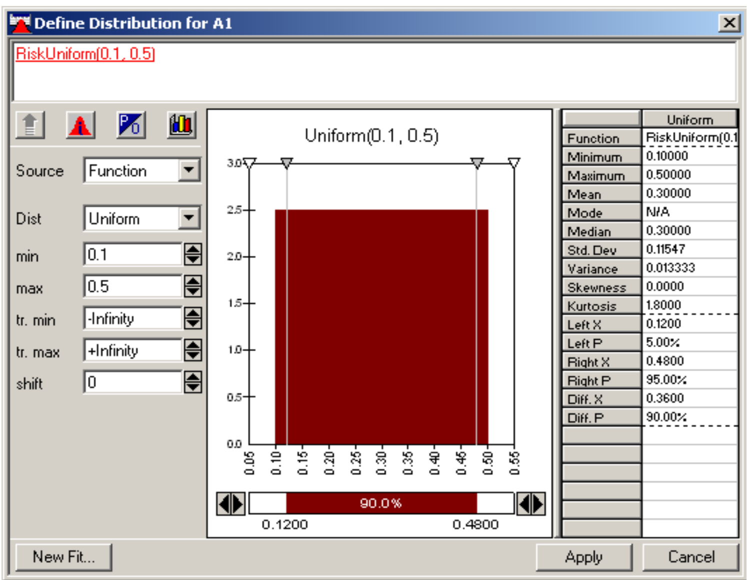

Figure I-6-1. Uniform Distribution using @Risk "Define Distribution"

If the answer to the question is yes, the facilitator will then ask, "what is the most likely value?" That value can then be used to define the peak of a triangular distribution, as shown in Figure 13-2. The mean of this distribution is simple the average of the three values (low, peak, high) that define the distribution.

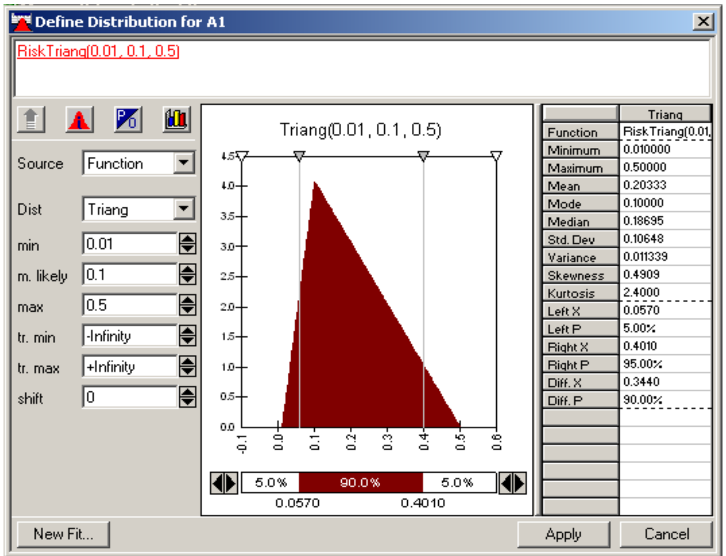

Figure I-6-2. Triangular Distribution using @Risk "Define Distribution"

Occasionally, there is sufficient information, that a more complicated distribution is appropriate. In this case, the questioning typically goes something like: "The probability is not likely to be less than _?_ (10th percentile) and is not likely to be more than _?_ (90th percentile). It cannot be less than _?_ (0 percentile) nor more than _?_ (100th percentile). It is equally likely to be more or less than _?_ (50th percentile). Using these data pairs, a cumulative probability distribution can be defined in @Risk, as shown in Figure 13-3.

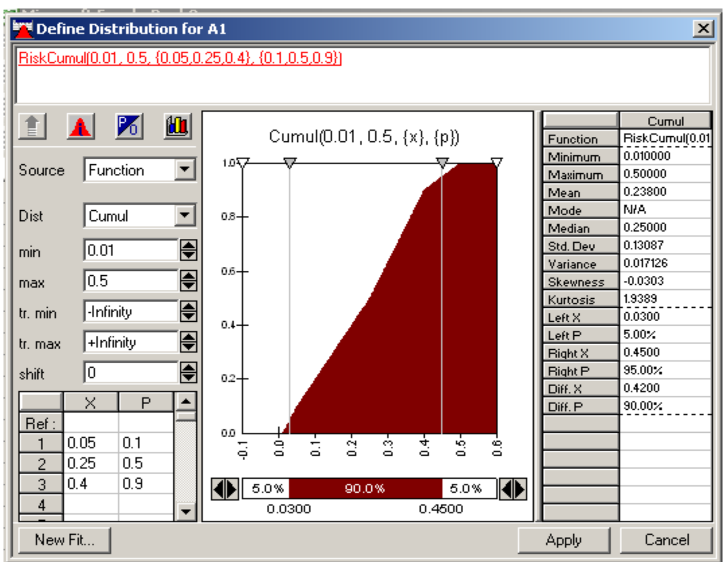

Figure I-6-3. Cumulative Distribution using @Risk "Define Distribution"

Biases

Biases can affect subjective probability estimates. These must be recognized, and to the extent possible, the risk analysis facilitator must strive to minimize their impacts to the estimates. Vick (2002) describes many of these in detail. Some of the more common are summarized briefly below.

- **Overconfidence Bias.** Perhaps the most pervasive bias is the tendency for people to be more confident than the evidence warrants. This usually leads to estimates that are closer to zero or to one than they really should be, and/or to distributions that are narrower than they should be. Unfortunately, the more expertise and knowledge one possesses, the greater the tendency for overconfidence bias appears to be. This tendency has been demonstrated in experiments involving the general population as well as engineers specifically. One way to help avoid this is to start with an exercise where estimators need to estimate the likelihood of known events, such as that given in the back of this section. People can then get a feel if they are overconfident.
- **Anchoring Bias.** This is the tendency for estimates to not vary much from values that are initially presented to the group, either from base-frequency information or other sources. This is not mentioned to discourage presentation of base-frequency information, which is often useful to arrive at a reasonable estimate. However, if this information is presented, it should be as appropriate as possible to the problem being estimated and consideration should be given as to how the specific case might vary from the general population. Anchoring bias can be avoided to some extent by discussing potential extreme values before settling on a "best estimate".
- **Availability Bias.** This results from over-emphasizing the most recent, most easily recalled, or most vivid evidence. For example, if the team just discussed a catastrophic failure case history, there could be a tendency for the team to assign a higher likelihood to that failure mechanism. This may be appropriate if relevant similar conditions exist between the failure case history and dam being studied, but may not be appropriate if the similarities are questionable. A facilitator can try to point out the relevance of the latest information and counter it with an opposing view, if it is not thought to be representative of the issue at hand.
- **Motivational Bias.** This results when one or more of the members making estimates have a vested interest in the outcome. For example, if the designer of the dam is in the room, that person may want to push the view that the dam is well designed and there should be no problems with it, and hence will provide low failure probability estimates. Or perhaps an office needs work, and their representative is quick to push for higher failure probability estimates. A facilitator can attempt to bring out opposing views when an opinion appears to be expressed from a motivational bias.
- **Representativeness Bias.** This results from overemphasizing similarities and neglecting other information. The probability of B, given A is not equal to the probability of A, given B. For example, if slides are observed on steep slopes, one might conclude that steep slopes cause slides, or you are likely to find slides in areas of steep slopes. In reality, it is telling us the opposite; that you are likely to find steep slopes in areas of sliding (Vick, 2002). While this distinction may seem unimportant, it can make a difference in the likelihood. The fact that slides occur infrequently on flat slopes might simply be that there aren't any flat slopes. We also don't know anything about the false positives, or steep slopes that are stable. This one is difficult to identify and correct, but if recognized by a facilitator, the appropriate questions can be asked (e.g. "what do we know about steep slopes that are stable?").

Group Dynamics

If the group selected to make the estimates has been chosen properly, people will enter the exercise with the appropriate expertise, an open mind, and a willingness to get to the best possible results. When this occurs, appropriate interactions take place, and additional ideas expand from those being discussed. Arguably, the group reaches a better conclusion than any of the individual members could have on their own. However, sometimes the group may stumble along one or more of the following lines. The facilitator must recognize when this is occurring, and try to direct the group toward a more positive direction.

- A dominant individual may drive the way the group goes by "bullying" everyone into thinking the way they do. It takes a fairly strong facilitator to deal with this, and usually requires emphasizing and bringing out the opposing point of view as well as drawing others into the conversation. Alternatively, an individual estimating procedure can be instituted to ensure everyone's input is received. **(Note: Individual estimating procedures can be instituted in other cases as well, but when using voting procedures, the group must decide how an expected value and distribution will be selected from the results.)**
- People may not say what they really feel for fear of appearing unknowledgeable, and will tend to go along with the rest of the group even though they have important input. This requires the facilitator to draw out their opinions by directing questions specifically at these individuals.
- The group gets tired due to the rigors of the session, and people agree just to get it done. The facilitator is not immune to this trap. If it is obvious that proper attention is not being paid to something, it is important to stop, take a break, and discuss ways to invest proper time for the evaluation.

An Iterative Probability Estimation Process is Desirable

Expert elicitation to obtain subjective probability estimates can be improved if an iterative process is used. During the expert elicitation process, discussion regarding a particular event attempts to reveal all possible evidence that would either make the event more or less likely. When the discussion tapers off, the facilitator employs a process to have each expert turn their considerations for the likelihood or truthfulness of the event into a numerical representation of a degree of belief. Within the USACE the facilitators will typically have the experts write their estimates on paper out of the sight of the other experts. Within Reclamation, the estimates are typically presented verbally. After the facilitator collects all estimates and makes them known to the entire estimating team, the estimates might all clump together about some common number, they might spread over a wide range, clump in two or more groups, or there might be a common group with one or two outliers. The facilitator would then call upon representatives of differing groups to explain why it was they held a particular belief in light of the evidence common to all. Discussion at this point would result in the facilitator asking if any members of the estimating group might care to change their estimate.

There are a number of values involved that address the reasons why estimators should or should not be allowed to change their estimates. The process should produce consistent results that would be repeatable if a different group of estimators were asked to assess the same event and evidence. Agreement between the estimators might indicate everyone is interpreting the information in the same way. Disagreement might indicate that some estimators are mistaken about the importance of particular evidence or that they hold different views in mind about geologic models or design or construction details. Disagreement might arise because some estimators may have a difficult time converting degree of belief to a numerical value. Accuracy or correctness is desirable, even if it only means that estimators are attempting draw from their experience to systematically capture judgment in a numerical estimate to the best of their ability. Any process that would introduce the various biases described elsewhere in this section is not desirable.

Allowing the team members to change their estimates affects most of these values in positive ways. When the differing groups are called upon to explain their viewpoints, the reasons why they differ become apparent. Impressions discovered to be mistaken should be corrected. Differing views of the world can arise from definitions that are unclear or from descriptions that are not sufficiently detailed. Additional discussion arising from the differing views can help to clarify these points and can help the estimators to evaluate similar conceptions. Difficulties some estimators might have in expressing judgment as a numerical probability estimate can be overcome by getting the estimator to use various verbal descriptors to qualify how frequently they would expect to see various occurrences.

Each facility's worst features, those that are most likely to end up as dam safety issues or deficiencies, are limited. A procedure to diagnose the issues, if used by different expert panels and if effective, should identify the same critical issues. The secondary discussion to correct mistaken impressions, to clarify definitions and better describe estimators' conceptual models, and to accurately represent individual's judgment all should all work to create estimates likely to be consistent or repeatable from one estimating team to another. The combined experience of several experts revealed by initial discussion to establish likely and unlikely factors and then revealed by discussion to understand differing interpretations will help to create consistent estimates.

There are reasons against having an iterative process for subjective probability estimation. Estimations are supposedly subject experts trained in elicitation techniques. If the first discussion round has been exhaustive, their first estimate might best represent their judgment. Allowing discussion after an initial round of estimates exposes the team members to the possibility of being affected by bias or by bullying. If one team member has a particularly strong personality they may try to force the others to their way of thinking. If one team member is highly respected, the other team members may get into the habit of listening to this individual and being swayed by their opinions. Availability bias suggests that the most recent information can be considered as most important. If lengthy discussion occurs over a particularly contentious issue, team members might focus on the evidence surrounding that issue allowing pertinent evidence surrounding other issues to weigh less heavily in their minds.

Overall, it is better to make the process iterative. The benefits of the additional discussion outweigh the problems that can arise. Foremost, when asked to state why they hold various opinions, the differing advocates will usually focus on the most important claims and evidence inherent to the controversy, which can be very helpful when it comes time to develop the dam safety case. Also, the rules of the subjective probability elicitation process dictate that if consensus cannot be achieved, both estimates are carried forward. Finally, the opposing viewpoints can often be turned into hypotheses which can be tested by additional investigations. A call for additional investigations can receive strong justification for funding when it can be shown that the results will reduce or eliminate disagreement over crucial issues. To reduce the potential for biasing or bullying, facilitators can be provided training to be able to recognize the problems and to be shown techniques to stop them.

Exercise

Self Test of Overconfidence

This test is adapted from Russo and Schoemaker (1990). Provide a low and high estimate for each of the following ten items such that you are 90 percent sure the answer lies between your estimates. Try not to be too overconfident (narrow range) or underconfident (wide range). If you miss more than one (ten percent), your knowledge is not as good as you think it is.

Item	90% Confidence Band	
	Low	High
Abraham Lincoln's age at death		
Length of the Nile River (in miles)		
Number of nations in NATO		
Number of studio albums released by the Beatles		
Diameter of the moon (in miles)		
Weight of an empty Boeing 747 (in pounds)		
Year in which Leonardo da Vinci was born		
Gestation period of an African Elephant (in days)		
Air distance from London to Sydney (in miles)		
Deepest known point in oceans (in feet)		

You might be thinking that it is possible for someone to be overconfident in a trivia test, but be good at estimating dam failure probabilities if that falls within their area of expertise. This could be the case, but tests involving employees of a chemical company, managers of a computer company, physicians, and physicists indicated many missed the mark to a similar extent as trivia tests even when asked questions about their own industry (Russo and Schoemaker, 1990).

References

Barneich, J., D. Majors, Y. Moriwaki, R. Kulkarni, and R. Davidson (1996), "Application of Reliability Analysis in the Environmental Impact Report (EIR) and Design of a Major Dam Project," *Geotechnical Special Publication No. 58*, American Society of Civil Engineers, New York, Volume 2, pp. 1367-1382.

Hartford, D.N.D. and Baecher, G.B. (2004), Risk and Uncertainty in Dam Safety, Thomas Telford Publishing, London, 391 p.

Reagan, R., F. Mosteller, and C. Youtz (1989), "Quantitative Meanings of Verbal Probability Expressions," *J. Applied Psychology*, vol. 74 no. 3, pp. 433-442.

Russo, J.E. and P.J.H. Schoemaker (1990), *Decision Traps*, *The Ten Barriers to Brilliant Decision-Making and How To Overcome Them,* Fireside/ Simon & Schuster, New York.

Vick, S.G. (2002), *Degrees of Belief, Subjective Probability and Engineering Judgment*, ASCE Press, American Society of Civil Engineers, Reston, Virginia, 455p.

Vick, S.G. (1999), Considerations for Estimating Structural Response Probabilities in Dam Safety Risk Analysis, Appendix A of Reclamation Dam Safety Risk Analysis Methodology (version 3.3), 27 p.
http://intra.usbr.gov/ssle/dam_safety/risk/Usbrmeth.newA.final.pdf

von Winterfeldt D. and W. *Edwards* (1986), *Decision Analysis and Behavioral Research*, Cambridge: Cambridge University Press. pp. 604.

This Page Left Blank Intentionally